Confidence in risk assessments

Jonathan Rougier*

School of Mathematics

University of Bristol

Abstract

Risk is assessed with varying degrees of confidence, and the degree of confidence is relevant to the risk manager. This paper proposes an operational framework for representing confidence, based on an expert's current beliefs about how her beliefs might be different in the future. Two modelling simplifications, 'no unknown unknowns' (NUU) and an homogeneous Poisson process (HPP) make this framework trivial to apply. This is illustrated for assessing an exceedance probability for a large event, with volcanic risk as a specific example. The paper ends with a discussion about risk and confidence assessments for national-scale risk assessment, including several further illustrations.

Keywords: Prospective interval, Reasonable Worst Case (RWC), exceedance probability, 'bazaar of experts', 'no unknown unknowns' (NUU)

*School of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW. Email: j.c.rougier@bristol.ac.uk.

1. Introduction

This paper is aimed at experts performing risk assessments, and the risk managers who commission them. I will assume just one expert, for convenience of expression, and imagine that she has been commissioned to provide a probabilistic risk assessment; more details of one such assessment are given below. The emphasis throughout is on practicality.

A risk assessment can be made with a varying degree of confidence, reflecting both the innate difficulty of the task, and the amount of resources available. The expert's confidence in her risk assessment is obviously relevant to the risk manager, and it is common for risk assessments to close with a question such as:

> How confident are you in your risk assessment?
>
> 1 = not confident at all, ..., 5 = highly confident.

However, this question contains a lot of reducible ambiguity. Statisticians strongly advocate the use of operationally-defined quantities when assessing uncertainty, so that the resulting assessment is not contaminated by ambiguity of meaning (Lad, 1996; Cooke, 2004). The presence of ambiguity makes it hard to interpret an answer, as illustrated in section 2.

Ambiguity also makes it harder to compare risk assessments across hazard classes, particularly if different hazard classes are assessed by experts coming from different risk cultures. The common need for a general-purpose approach which can be applied across hazard classes also rules out technically sophisticated approaches such as intervals, e.g. using lower and upper previsions (Troffaes and de Cooman, 2014). However attractive these might be to mathematicians, it is hard to disambiguate such intervals at an acceptable cost, for use by experts working across a range of fields.

I propose to replace the ambiguous question above by a more operationally-

defined question, adjusted to the particular needs of the risk manager. As an example:

> Imagine performing this risk assessment again in five years' time. What is your current 90% uncertainty interval for the value you will assign in five years' time?
>
> Lower: ☐ , Upper: ☐ .

I think most people would intuit that a '90% uncertainty interval' will contain the specified value in about 90% of such assessments. This can be made precise with written guidance, if necessary.

Part of the idea of this proposal is to put cognitive roadblocks in place, to prevent a facile answer—what Daniel Kahneman would call a 'system one' answer (Kahneman, 2011). The process of risk assessment is exhausting, as is the process of filling-in a risk assessment questionnaire. If by the end of the process the expert is exhausted, her answers will tend to be less considered. So a simple question at this stage can provoke a facile answer, particularly if it stands between the expert and a cold beer. A more complicated question, which needs to engage 'system two' to process it, is presumably more likely to get a 'system two' response.

But there are other advantages as well. First, the operational nature of the question means that the outcome is verifiable, discussed further in section 5. Second, the question can be aligned with the needs of the risk manager, who may be able to defer a decision if the information acquired in the near future might change his choice of action. Thus the risk manager specifies the delay for each hazard class, according to his timetable and his priorities: section 2 has an illustration.

Third, the question can be answered by direct calculation, given a statistical model. This is the topic of most of this paper, sections 3 and 4, 6 and 7, and illustrated in sections 8 and 10. Suffice to say that the expert does not need a statistical model, but she may find it helpful to reflect on the output of

a statistical model, even quite a simple one, as discussed further in sections 9 and 10; this final section considers national-scale risk assessment.

2. Intermission: 'Soft' and 'hard' probabilities

This section, about a fictional treatment of an actual event, is primarily a warning against making complicated decisions in testosterone-filled environments. But it also illustrates the difficulty in interpreting ambigiuous assessments of probability and confidence, and how they might be resolved.

Readers may recollect the film Zero Dark Thirty (2012, directed by Kathryn Bigelow). In a pivotal scene, the Director of the CIA goes around the table asking each person for their probability that the unknown man in the compound in Abbottabad, Pakistan, is Osama Bin Laden. One of them (Daniel) replies "I'd say it's a soft sixty, sir. I'm virtually certain there's some high value target there, I'm just not sure it's Bin Laden." Shortly after, Maya, the analyst, who is sitting at the back, gets frustrated and blurts out "One hundred percent, he's there; okay, fine, ninety-five percent because I know certainty freaks you guys out; but it's a hundred!".[1]

We put aside that these are fictional characters, to explore what they mean. Daniel is virtually certain about one proposition, 'the man is a high-value target'. He does not indicate his confidence but we may assume, from the fact that he does not feel the need, that he is confident. He is less certain about a second proposition 'the man is Bin Laden', $p = 60\%$, and this he qualifies with the adjective 'soft', indicating a lack of confidence. Maya has $p > 95\%$ for this second proposition, and uses her demeanour to signal confidence: her probability is a 'hard' probability. In two lines of dialogue, these two experts have apparently delivered a large amount of information to the risk manager

[1] Taken from the script at `http://www.imsdb.com/scripts/Zero-Dark-Thirty.html`. Readers of a sensitive disposition should be warned that the language is strong in this testosterone-fuelled scene, including in a further quote below, in which the expletives have been replaced by ellipses.

(the Director), covering both their probabilities and their confidences.

Or have they? Unfortunately, 'soft' and 'hard' are also capable of a second interpretation in this dialogue. Maybe by 'soft' Daniel means 'at most 60%', and by her demeanour Maya means 'at least 95%'. We will never know, and neither will the Director, because he does not ask for clarification. Instead of accepting this ambiguity, he could have asked for confidence in a much more specific and useful form. He has a window in which he can launch a mission: perhaps three weeks. What he really wants to know, it seems to me, is how much each person thinks their current probability might change in the light of another week or two of intelligence. In other words, he wants something like a probability and a 90% 2-week uncertainty interval.

Of course it is not necessary to be so precise, either about the level or about the delay. In the film, he instructs his experts: "I'm about to go look the President in the eye and what I'd like to know ... is where everyone stands on this thing. Now, very simply. Is he there or is he not ... there?". Instead, he could have asked "I want each of you to give me your probability that he's there, and also tell me whether you think your probability might change meaningfully if we acquire another week or two of intel." And he might have added, "To keep things snappy, address the second question on a scale from 'soft' to 'hard', where 'soft' indicates meaningful change, and 'hard' indicates no meaningful change." Then Daniel's 'soft sixty' and Maya's 'hard ninety-five' would have been less ambiguous and more useful.

At the risk of further dulling the excitement of the scene, I would disambiguate 'meaningful'. The Director is reluctant to launch a mission unless it has a high probability of success. So he wants to know whether there is a reasonable chance that the probability of success will drop below some threshold, say 50%. There is a sizable probability that even if the man in the compound is Bin Laden, the mission will fail, as nearly happened in the film. If there is a 20% probability of failure then he needs the probability of the man being Bin

Laden to be at least $0.5/0.8 \approx 60\%$. However, I would advise him not to reveal his threshold of 60% before the answers, because of the danger of 'anchoring' (Kahneman, 2011). He would be best served by asking for an actual interval of probabilities, rather than a qualitative label like 'soft' or 'hard'. But if that requires too much of his experts, he still extracts more useful information by tying 'soft' and 'hard' to an uncertainty interval with a specific delay, than he does by not asking about confidence and misinterpreting an ambiguous response.

3. Risk and the 'reasonable worst case' (RWC)

In a given hazard class, there will be a range of possible occurrences varying primarily by magnitude, but also by other features such as intensity and duration. For simplicity, I will take magnitude as a scalar proxy for the type of occurrence. Each occurrence generates a loss. For simplicity, I will take financial loss as a scalar proxy for the aggregation of the various dimensions of loss. Taken together, (time, magnitude, loss) comprises a 3D point process Π say, with domain $(0,1) \times \mathbb{R}_+ \times \mathbb{R}_+$ over the forthcoming year. Risk assessment concerns the distribution of the annual loss,

$$A := \sum_{(T,M,L) \in \Pi} L, \tag{1}$$

where T is time, M magnitude, and L loss.

The process Π is very complicated to assess, and it is natural to ask whether there are some reasonable conditions under which key features of A can be assessed using a much simpler construction. One feature which is amenable in this way is the expected annual loss. If there is a 'Goldilocks' magnitude $\widetilde{m}$, sufficiently small that $\mathbb{E}(L \mid M \leq \widetilde{m}) \approx 0$, and also sufficiently large that $\mathbb{P}(\widetilde{N} > 1) \approx 0$, where $\widetilde{N}$ is the number of occurrences with $M > \widetilde{m}$ in the

next year, then

$$\mathbb{E}(A) \approx \mathbb{P}(\widetilde{N} > 0) \cdot \mathbb{E}(L \mid M > \widetilde{m}), \tag{2}$$

by the Law of Iterated Expectation. Eq. (2) is a formal basis for the mantra 'risk is probability times consequence', where risk is taken to be expected annual loss. $\mathbb{P}(\widetilde{N} > 0)$ is termed the 'exceedance probability' of $\widetilde{m}$: the probability of at least one occurrence of $M > \widetilde{m}$ in the next year.

The Goldilocks condition seems strong, but it can be a reasonable approximation when loss is a strongly convex function of magnitude. In this case, the expectation of A will be dominated by low-probability high-magnitude occurrences, the losses from which will far exceed the losses from low- and medium-magnitude events which will be, relatively speaking, negligible. This seems to me to be the best justification for the very common practice of representing a hazard class by a single large-magnitude occurrence $\widetilde{m}$, and computing risk of the hazard class as the product of the exceedance probability of $\widetilde{m}$, and a value for the loss should there be an occurrence with $M > \widetilde{m}$.

This practice is ubiquitous in national risk assessement, discussed in section 10. In the UK National Risk Assessment (NRA), for example, each hazard class is represented by a 'reasonable worst case' scenario (RWC, see NRR, 2017). Each hazard class is plotted on a common risk matrix as a ('likelihood', 'impact') dot for the RWC. Under my interpretation of the NRA, the Goldilocks condition is assumed for each hazard class, and the RWC is identified with a specified magnitude $\widetilde{m}$. The 'likelihood' of the hazard class is the exceedance probability $\mathbb{P}(\widetilde{N} > 0)$, and the 'impact' of the hazard class is the expected loss $\mathbb{E}(L \mid M > \widetilde{m})$. The NRA focuses on the hazard classes in the top righthand corner of the risk matrix. Under my interpretation this is correct, as these hazard classes have the largest expected annual loss, i.e. the largest risk.

Now accept the Goldilocks condition, for some specified magnitude $\widetilde{m}$. Of

the two quantities $\mathbb{P}(\widetilde{N} > 0)$ and $\mathbb{E}(L \mid M > \widetilde{m})$, the former is usually much more challenging than the latter. In fact it is quite common to narrate the RWC event, or to choose it from among well-documented historical events. For example, the Carrington event (1859) is a useful RWC scenario for a massive space weather event. The RWC narrative fixes $\widetilde{m}$ and allows a fairly straightforward assessment of $\mathbb{E}(L \mid M > \widetilde{m})$. Therefore, the hard part of risk assessment is specifying the exceedance probability $\mathbb{P}(\widetilde{N} > 0)$. In the rest of this paper, I will focus on specifying the exceedance probability for some stated threshold. Technically, this would be an 'uncertainty assessment' rather than a 'risk assessment', but, as I have explained, in the context of the Goldilocks condition this distinction is blurred.

4. Beliefs, through time

Risk assessment is about judgement, and judgement is a property of the mind; i.e., it is 'personal' to use the preferred term of the great 20th century statistician L.J. Savage (see, e.g., Savage, 1971). The risk manager, or his auditors, cannot require of the expert that she give a traceable account of precisely how she arrived at her judgements. This is unattainable if those judgements have been gradually constructed over decades of study, discussion, experiment, analysis, and reflection. What the risk manager requires is that the expert is indeed an expert for the hazard class in question, and that she has honestly reflected her judgement in her risk assessment: i.e., she accepts ownership of her judgements. It is common to use 'judgement' when referring to an expert's uncertainty (Aspinall and Cooke, 2013), but below I will use 'belief', because of its presence in philosophers' quick definition of 'knowledge', which is 'justified true belief' (Ladyman, 2002, pp. 5–6). To me, 'judgement' seems a bit ponderous, and liable to be misconstrued.

Our beliefs are always 'time-stamped': we hold them at a point in time,

conscious that they have changed in the past, and may change in the future. The most obvious cause of a change is the availability of new data. To change our beliefs substantially, these data will usually be anomalous. For example, the tsunami following the Tōhoku earthquake in Japan in 2011, or the Kaikoura earthquake in New Zealand in 2016, both of which were much larger than expected; or the Grenfell Tower fire in the UK in 2017, which revealed compliance issues with building regulations. But it is also possible that the existing data, or the theory within which they are interpreted, will be reappraised. Part of a person's status as an expert in her field is that she has useful beliefs about these possibilities. These beliefs will be reflected in her confidence assessment.

For concreteness, suppose that the expert is required to provide her exceedance probability for a 'reasonable worst case' occurrence in some hazard class (see section 3). This is her probability of at least one occurrence at least as bad as the reasonable worst case occurrence happening in the next year. Rather than write 'at least as bad as the reasonable worst case occurrence' I will just write 'large occurrence'. So the exceedance probability is the probability of at least one large occurrence happening in the next year.

Let N be a counting process, such that $N(t, t')$ is the number of large occurrences happening in the time interval (t, t'). Let $t = 0$ denote the present, and $k > 0$ denote some specified time k years into the future. Then

$$p_0 := \mathbb{P}_0\{N(0,1) > 0\} \tag{3a}$$

denotes the expert's current exceedance probability, and

$$P_k := \mathbb{P}_k\{N(k,k+1) > 0\} \tag{3b}$$

denotes her exceedance probability in k years time. P_k is a random quantity

at time $t = 0$, with distribution function

$$F_{0,k}(u) := \mathbb{P}_0(P_k \leq u). \tag{4}$$

An expert's current uncertainty about P_k can be captured in a *prospective interval*, constructed from $F^-_{0,k}$, the quantile function of $F_{0,k}$:

$$F^-_{0,k}(p) := \inf\{u \in \mathbb{R} : F_{0,k}(u) \geq p\}. \tag{5}$$

Here is a specific definition, to avoid more notation.

Definition 1 (Prospective Interval)**.** Your 90% 5-year prospective interval for a proposition A comprises the 5th and 95th percentiles of your current probability distribution for the probability you will assign to the analogue of A in 5 years time. In the notation of this section, $F^-_{0,5}(0.05)$ and $F^-_{0,5}(0.95)$.

The notation in this section will be useful below, but its initial purpose is to give a precise form to the idea that the expert can be currently uncertain about her future probabilities, and can—in principle—capture her current uncertainty in terms of a distribution function. In these terms, the proposed confidence question from section 1 is to ask the expert for her lower and upper limits of some appropriate prospective interval. The choice '90% 5-year' is a generic suggestion, to be modified in particular applications. For terrorism risk, for example, a delay of 5 years is likely to be too long (Woo, 2015).

This concept of currently uncertain future probabilities has a respectable provenance in both Statistics (Goldstein, 1997, 'temporal sure preference'), and in Philosophy (van Fraassen, 1995, 'reflection principle'). In these two treatments the event is the same, but in my treatment the event may have to change, although the change can be subtle. In section 2, the event under

discussion in the boardroom is:

$$A = \text{'The man currently in the compound is Bin Laden'}.$$

Returning to the boardroom in two weeks time, the 'currently' in A will be two weeks advanced, and so the analogous event in the prospective interval is slightly different. For exceedance probabilities, the current exceedance probability is for 2020, say, but the 5-year prospective interval exceedance probability is for 2025. The degree to which a prospective interval provides a confidence assessment of a current probability depends on the closeness of the analogue. As the illustrations in this paper demonstrate, there are plenty of close analogues in risk assessment.

5. The 'bazaar of experts'

From the expert's point of view, a single probability p_0 is already more of a hostage to fortune than a verbal label such as 'quite unlikely'. Providing a 90% five-year prospective interval (ℓ, u) raises the stakes much higher. In five years time, the expert will be aware of whether her p_5 lies inside her original (ℓ, u). If it is not inside, she faces a tricky decision from a personal point of view, about whether to reveal her true p_5 and the deficiency of her original (ℓ, u), or whether to adjust her p_5 towards/into her original (ℓ, u). In the former case her reputation suffers, while in the later case she may appear out of step with other experts, many of whom are not constrained by previously-stated prospective intervals. On the other hand, she might be able to celebrate the fact that her true p_5 is inside her original (ℓ, u), in which case her reputation is enhanced.

The logic of this situation will be clear to both risk managers and experts. An expert who provides a prospective interval puts more of her reputation on the line; gamblers would say she has more 'skin in the game'. Thus she

distinguishes herself from other experts by signalling her higher level of commitment. This is more than a binary signal. First, her prospective interval can be narrower than others'. Second, she can make more predictions and provide more prospective intervals: for example, she might provide prospective intervals for different durations, or over a range of events within her expertise.

In a world scarce of experts, this type of signalling would be unnecessary. But we live in a world full of 'experts', and the challenge for the risk manager is to select a good one (Tetlock, 2005). In this 'bazaar of experts', committed experts will want to distinguish themselves by providing a benchmark by which they can be judged. In so doing, these experts provide a defensible reason for risk managers to select them over other experts who are unable or unwilling to quantify their confidence in their predictions. In turn, risk managers can justify their choice of experts to their clients' auditors (Smith, 2010).

Finally, it is worth mentioning a valuable tool for guarding against overconfidence—represented in this context by a too-narrow prospective interval. This is the 'premortem', originally proposed by Gary Klein (Klein, 2007), and discussed by Daniel Kahneman (Kahneman, 2011, p. 264). Adapted to this context:

> Imagine we are five years into the future. Owing to events in the last five years your new probability is substantially higher than the upper limit of your original 90% 5-year prospective interval. Take 5 to 10 minutes to write a brief history of those events.

I have used 'higher', but 'lower' or 'outside' might be more appropriate, depending on circumstances.

6. No unknown unknowns (NUU)

A current distribution function for a future probability is a complicated thing. In this section and the next I show how it can be generated by a simple algo-

rithm under two modelling conditions, 'no unknown unknowns' (this section) and a homogeneous Poisson process (section 7). I will continue to use the example from the previous section, where p_0 is the expert's current exceedance probability, and P_k is her exceedance probability k years into the future, which is currently an uncertain quantity.

As already indicated, there need be no connection at all between p_0 and P_k, because of all the things that might happen in the interval $(0, k)$, some of which might be surprising to the expert. But there ought to be a relationship between p_0 and $F_{0,k}$, because both are based on the same information—the information available to the expert at time $t = 0$. This section presents a probabilistic model in which this relationship is explicit, which I term 'no unknown unknowns' or 'NUU'. NUU is descriptive of the expert's current state of mind. That is, the expert examines her beliefs at time $t = 0$ and decides that they are consistent with NUU, and is then able to apply NUU to compute her p_0 and $F_{0,k}$.

Definition 2 (No unknown unknowns, NUU)**.** The modelling framework of NUU comprises:

1. A specified stochastic process generating N, starting at time $-a < 0$ (i.e. a years in the past), and

2. Probability at time t is computed by conditioning on the outcome of the stochastic process over the interval $(-a, t)$.

Updating probabilistic beliefs by conditioning is known as 'Bayesian conditionalization' in the philosophy of inference (Howson and Urbach, 2006, ch. 3). According to NUU, an 'unknown unknown' is belief-changing information which cannot or should not be conditioned upon. This, I propose, is the mathematical representation of Donald Rumsfeld's famous 'unknown unknowns' utterance: unknown unknowns change the expert's stochastic process.

To avoid any more notation, I will conflate the history of the process with the process itself, because the generalization is straightforward. The stochastic process is $\mathbb{P}_{-a}$. Under NUU,

$$\mathbb{P}_0(\cdot) = \mathbb{P}_{-a}\{\cdot \mid N(-a, 0) = m\}, \tag{6a}$$

where m is the known number of large occurrences which have happened over the interval $(-a, 0)$. Then p_0 from (3a) has the form

$$p_0 = \mathbb{P}_0\{N(0, 1) > 0\}. \tag{6b}$$

Similarly, P_k from (3b) has the form

$$P_k(y) = \mathbb{P}_0\{N(k, k+1) > 0 \mid N(0, k) = y\}, \tag{6c}$$

based on y occurrences in the interval $(0, k)$. From the point of view of $t = 0$, though, y is uncertain, and hence $F_{0,k}$ from (4) has the form

$$F_{0,k}(u) = \mathbb{P}_0\{P_k(Y) \leq u\}, \tag{6d}$$

writing $Y := N(0, k)$. This function can be expressed algorithmically as

$$F_{0,k}(u) = \sum_{y: P_k(y) \leq u} \mathbb{P}_0(Y = y). \tag{6e}$$

Eq. (6) are the NUU equations.

The effect of NUU is to make all current and future beliefs a function of the stochastic process $\mathbb{P}_{-a}$ and the history over $(-a, 0)$. But NUU does not impose any simplifications on the expert's stochastic process $\mathbb{P}_{-a}$, which can be as rich and as complex as she requires. The sequence of large occurrences in the time interval $(0, k)$ will certainly be in P_k, but so will the expert's beliefs

about the response to those occurrences, if she chooses. For example, with industrial hazards, she might believe that an occurrence of a certain type will cause the regulator to shut down or modify some installations, which will then reduce the probability of further occurrences of that type to near-zero, and so on. In this way NUU provides a licence for the expert to specify a rich $\mathbb{P}_{-a}$ because her p_0 and $F_{0,k}$ can be extracted algorithmically using (6).

7. NUU with an homogeneous Poisson process (NUU-HPP)

Having just extolled the virtues of a complex stochastic process, this section presents the NUU calculation of p_0 and $F_{0,k}$ based on the very simplest interesting stochastic process, an homogeneous Poisson process (HPP) (see, e.g., Kingman, 1993; Davison, 2003). I will term this combination 'NUU-HPP'. The expert decides to treat N as an HPP with unknown rate λ. This is how the expert chooses to model her beliefs, not in any sense a statement about nature itself. The following probability theory is standard (see, e.g., Davison, 2003; Robert, 2007).

Under the HPP model, the expert's beliefs are

$$\mathbb{P}_0\{N(t,t') = y \mid \lambda\} = e^{-(t'-t)\,\lambda}\,\frac{\{(t'-t)\,\lambda\}^y}{y!}, \quad y = 0, 1, \ldots . \tag{7}$$

The Jeffreys prior for λ is $\pi(\lambda) \propto \lambda^{-\frac{1}{2}}$, which is a natural choice here, and has the distinct advantage that all expressions are all available in closed-form. This prior can be embedded within the Gamma distribution as $\pi(\lambda) = \text{Gamma}(\frac{1}{2}, 0+)$, where $\frac{1}{2}$ is the shape parameter and $0+$ is the rate parameter; below, I use α and β, respectively. '0+' is a value a tiny bit larger than 0, a minor contrivance to keep π proper. Then, at time $t = 0$, having seen m large events in

the interval $(-a, 0)$, the conditional distribution of λ is

$$\pi_0(\lambda \mid m) := \text{Gamma}(\tfrac{1}{2} + m, a) \tag{8}$$

using the standard Poisson-Gamma conjugate update.

To evaluate p_0,

$$\begin{aligned} p_0 &= \mathbb{P}_0\{N(0,1) > 0\} \\ &= \int \mathbb{P}_0\{N(0,1) > 0 \mid \lambda\}\, \pi_0(\lambda \mid m)\, \mathrm{d}\lambda \\ &= \int (1 - e^{-\lambda})\, \pi_0(\lambda \mid m)\, \mathrm{d}\lambda \\ &= 1 - \left(\frac{\beta_0}{1+\beta_0}\right)^{\alpha_0}, \end{aligned} \tag{9}$$

where $\alpha_0 := \frac{1}{2} + m$ and $\beta_0 := a$, from (8). Eq. (9) may seem unfamiliar, but for large a, $p_0 \approx (\frac{1}{2} + m)/(1 + a)$ using a first-order approximation from the generalized Binomial theorem, which will typically be close to m/a, as expected.

Exactly the same reasoning applies to evaluate $p_k(y)$, but with

$$\pi_k(\lambda \mid m + y) := \text{Gamma}(\tfrac{1}{2} + m + y, a + k) \tag{10}$$

instead of $\pi_0(\lambda \mid m)$:

$$P_k(y) = 1 - \left(\frac{\beta_k}{1+\beta_k}\right)^{\alpha_k(y)}, \tag{11}$$

where $\alpha_k(y) := \frac{1}{2} + m + y$ and $\beta_k := a + k$, from (10).

Finally, to evaluate $\mathbb{P}_0(Y = y)$,

$$\begin{aligned} \mathbb{P}_0(Y = y) &= \int e^{-k\lambda} \frac{(k\lambda)^y}{y!}\, \pi_0(\lambda \mid m)\, \mathrm{d}\lambda \\ &= \frac{\Gamma(\alpha_0 + y)}{\Gamma(\alpha_0)\, y!} \left(\frac{k}{\beta_0 + k}\right)^y \left(\frac{\beta_0}{\beta_0 + k}\right)^{\alpha_0}, \quad y = 0, 1, \ldots, \end{aligned} \tag{12}$$

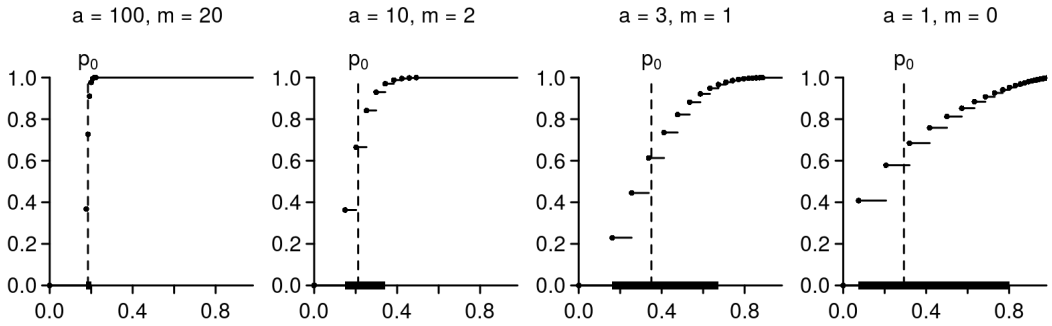

Figure 1: The NUU-HPP model: no unknown unknowns and an homogeneous Poisson process. Values for p_0, $F_{0,5}$ and the 90% 5-year prospective interval, for different values of a and $m = \texttt{round}(0.2 \cdot a)$. p_0 is shown as the vertical dashed line. $F_{0,5}$ is shown as the dots and horizontal bars. The prospective interval is shown on the horizontal axis as a thick bar.

where Γ denotes the Gamma function. Eq. (12) is a Negative Binomial distribution in an unfamiliar parameterization. The familiar parameterization has $\texttt{size} = \alpha_0$ and $\texttt{prob} = \beta_0/(\beta_0+k) = \beta_0/\beta_k$, according to the arguments of the `dnbinom` function in the statistical computing environment R (R Core Team, 2017).

Together, (9), (11) and (12) make up (6) under the NUU-HPP model. R code for computing p_0, $F_{0,k}$ and $F^-_{0,k}$ as functions of a, m, and k is given in the Appendix.

Figure 1 shows the NUU-HPP model in action, for different values of a and m, with $k = 5$. The overall pattern is very intuitive. When a (the history) is long and k (the future) is short, the expert's current uncertainty about P_k is small, because the information gained about λ over the interval $(0, k)$ is only a small addition to the large amount of information already gained over $(-a, 0)$. As a shortens, so her current uncertainty about P_k increases. When a is reduced to 1 year, her 90% 5-year prospective interval is roughly $(0.1, 0.8)$. At this point, she is close to declaring that she currently has no idea about what her future exceedance probability will be. An expert claiming high confidence in this situation cannot be using the NUU-HPP model.

Not a confidence interval. The definition of a prospective interval (section 4) makes it clear that a prospective interval is not a confidence interval (as defined, for example, in Casella and Berger, 2002, ch. 9). The NUU-HPP model, though, is a parametric model in which the parameter λ translates in a simple fashion into an exceedence probability:

$$\mathbb{P}\{N(0,1) > 0; \lambda\} = 1 - \exp(-\lambda) \approx \lambda, \tag{13}$$

the approximation holding for $\lambda \ll 1$. Therefore, a 90% confidence interval for λ can easily be transformed into a 90% confidence interval for the exceedance probability.

However, we should protest if the expert claims, as she may be tempted to do, that such a confidence interval is a good way to quantify 'confidence' in a manner suitable for a risk manager and his client. The deficiencies of confidence procedures in this regard are well-known (Morey et al., 2016). But it is also clear that prospective intervals have a decision-relevant control parameter 'k-year' which is absent from confidence intervals. It is a intrinsic feature of risk management to care about time horizons, but a confidence interval has no capacity to distinguish between waiting for 2 weeks, as in the Zero Dark Thirty scenario (section 2), a year, or five years or longer, as might be suitable for natural hazards.

8. Example: Icelandic volcanic risk

Consider the case of Icelandic volcanic risk, for which the reasonable worst case occurrence might be an explosive eruption of magnitude $M \geq 4$, i.e. at least 100 Mt of ejected matter (see, e.g., Sparks et al., 2013). In this section I assess my probability of exceedance, i.e. the probability that there will be at least one explosive $M \geq 4$ eruption in Iceland in the next year.

I believe that there are unlikely to be any volcanic 'unknown unknowns'

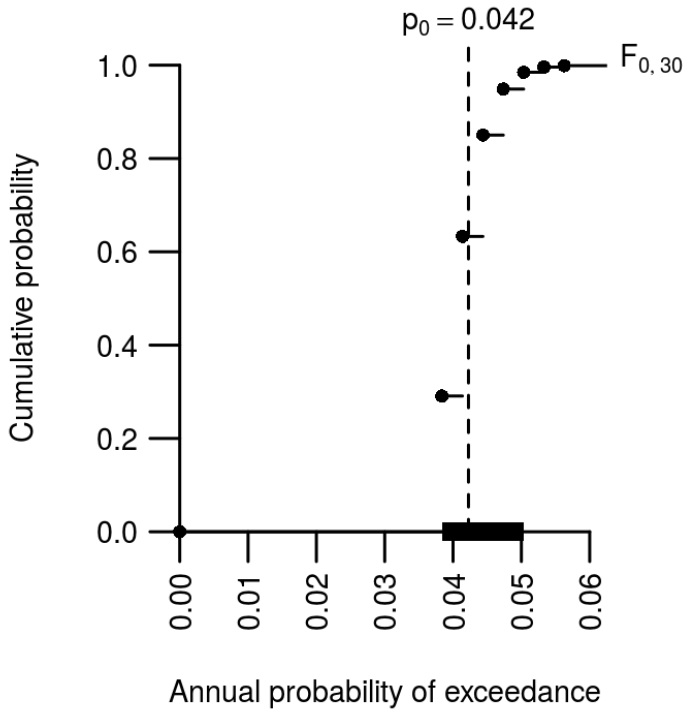

Figure 2: Icelandic volcanic risk, under the NUU-HPP model. According to the LaMEVE database, there have been 12 large ($M \geq 4$) explosive eruptions in the last 289 years, which gives an annual exceedance probability of $p_0 = 0.042$, with a 90% 30-year prospective interval of $(0.038, 0.050)$.

arising in Iceland the next few years, and adopt NUU. I believe that an HPP is a suitable model for large eruptions of groups of volcanoes for periods of at least a thousand years (Rougier et al., 2016). I will use the LaMEVE database to assess a and m (Crosweller et al., 2012; Brown et al., 2014), version dated 7 June 2018. As discussed in Rougier et al. (2018a), two relatively recent events create gaps in the Icelandic record: the shipwreck of Hannes Þorleifsson in the late 17th century, and the 1728 fire in Copenhagen. Therefore I use the LaMEVE record since 1730. Relative to today (end of 2018, there were no large events in 2018), this gives $a = 289\,\text{yrs}$, and $m = 12$. Figure 2 shows the resulting assessment: $p_0 = 0.042$ with a 90% 30-year prospective interval of 0.038 to 0.050. The 90% *5-year* prospective interval is 0.042 to 0.045, but a longer duration is often more appropriate for a natural hazard.

I am not an expert on Icelandic volcanoes. Like everyone else, though, I am

entitled to form my own beliefs, and sometimes to hold them strongly. It is the responsibility of the risk manager to choose his expert carefully. In the 'bazaar of experts' (see section 5) I hope the risk manager would be impressed by the transparency of my reasoning, and my willingness to provide a benchmark by which I can be judged. But I would fully expect him to select a volcanologist who is able to provide both a probability and a personal confidence assessment, and whose reasoning is at least as transparent as mine. My role as a 'statistician expert' is to set the bar on what is required of a 'subject matter expert'.

9. Richer models, and a caveat

NUU-HPP is the simplest case of a class of models which might be used by the expert within the NUU framework to model her beliefs. Under NUU, p_0 and $F_{0,k}$ are always computed using (6), but the stochastic process $\mathbb{P}_{-a}$ can be much richer.

One direction is to allow the process for N to remain Poisson, but to be non-homogeneous. For example, λ could be a known function of time, or a known function of other known variables, or an unknown function of those variables. If an unknown function, then $\mathbb{P}_{-a}$ would need to include beliefs about the functional form (e.g. in terms of a finite set of parameters). Those other variables could be known in the future or unknown in the future. If they are unknown in the future, then $\mathbb{P}_{-a}$ would need to include beliefs about those variables in the future. And so on.

Another direction is to move away from a Poisson process for N, to allow the process to be self-exciting or self-damping. For example, a self-exciting process might be used to model terrorism events, under the belief that successful events spawn new events. Although this has to be set against the post-event response of the security services to reduce the success probability of similar

events in the future. I am not an expert in this field, and my speculations are not worth much.

A *caveat.* I would caution against using the richer models described here. Each extension of HPP requires a substantial increase in the amount of statistical modelling, difficulties in selecting a prior distribution, and extensive computation, all of which are absent in the HPP case of a single unknown λ.

Statistical models are blunt tools, and we do not expect to represent all of our relevant beliefs in them (Cox and Donnelly, 2011). Instead, we use them as a springboard. The point about NUU-HPP is not that it is an adequate reflection of our beliefs, but that it is very easy to use. The expert who wants to compute a p_0 and a 90% 5-year prospective interval only has to ask herself a single question: how much of the well-recorded past do I believe is relevant for the next 5 years? Once she has answered this question with her choice of a, and found m from the database, she has candidate values for p_0 and her prospective interval. She could work a lot longer and harder to develop a more general model, making her values hostage to her modelling assumptions and coding, and still not close the gap very much between her statistical model and her beliefs. Better to use the NUU-HPP model, report its results, and then decide whether to let them stand, or to adjust them in some way. The following final section puts this suggestion into a more formal framework.

10. Implications for national-scale risk assessment

> "The database is too small for confident predictions."
>
> (Perrow, 2007, p. xxvi)

Most OECD countries undertake a periodic national risk assessment (OECD, 2018), which is required to evaluate existing response and recovery capabilities, and consequently to shape spending priorities. Confidence measures are valu-

able to risk managers (senior civil servants and government ministers) because they draw attention to the high-risk low-confidence hazards which complicate policy decisions. Such hazards could be the target of science research funding.

National-scale hazards span many classes, covering natural hazards (e.g., volcanic eruptions), biological hazards (flu epidemic), accidents (pollution of the water supply), and malicious acts (terrorism). The purpose of national risk assessment is to compare hazard classes, and therefore it is important that all hazard classes be assessed using the same metrics. This is very challenging, given that different hazard classes have very different profiles in terms of their impact, and that the risk cultures of subject matter experts (SMEs) in different hazard classes can vary widely. Another challenge is that risk assessments must be regularly updated, given that the national risk landscape is constantly changing. These challenges favour simple and transparent methods which can be widely applied, over bespoke methods that emerge from the deliberations of SMEs separately for each risk class.

The evidence suggests that people make better uncertainty assessments when they anchor on an empirical base-rate (Tetlock and Gardner, 2015). With this in mind, and favouring simplicity and transparency, I propose the following approach, for each hazard class:

1. Compile or identify a publicly-available database of occurrences of the hazard class.

2. Identify the starting-point of the relevant history in that database; this gives a value a for the length in years of the relevant history and m the number of large occurrences in that history.

3. Use a and m within a NUU-HPP framework to produce an exceedance probability p_0 and a confidence assessment (ℓ, u) for some specified time into the future, e.g. a 90% 5-year prospective interval.

4. Publish the source of the data, the values a, m, p_0, and (ℓ, u), and an explanation of why a was chosen.

5. Provide final values for probability and confidence, plus an explanation, if required, of why these differ from the database values.

The explanations in the final stage might not be publicly-available, if restricted information is used, e.g. in the case of terrorist acts.

Across hazard classes, the biggest source of additional relevant information is likely to be near-misses (Woo, 2018). For example, globally there have been 3 'super-eruptions' in the last 100 thousand years, for an exceedance probability of $p_0 = 3.5 \times 10^{-5}$, with a 90% 30-year prospective interval of nearly zero width. A more nuanced approach also considers eruptions smaller than super-eruptions, which can be incorporated into the exceedance probability of super-eruptions by assuming a smooth magnitude-frequency curve. This raises the exceedance probability of super-eruptions to 5.9×10^{-5} (Rougier et al., 2018b).

Some hazard classes will need rescaling to represent risk at the national scale, which will vary by nation. For example, there have been three major accidents at nuclear power facilities in approximately 17 thousand reactor-years of commercial nuclear power operation, according to `http://www.world-nuclear.org/information-library/safety-and-security/safety-of-plants/safety-of-nuclear-p` `aspx` (downloaded 28 Jan 2019). Based on these figures, my exceedance probability for a major accident at a nuclear reactor is $p_0 = 0.21 \times 10^{-3}$ with a 30-year prospective interval of nearly zero width. The UK Office for Nuclear Regulation (ONR) regulates 36 nuclear facilities, and ignoring, for simplicity, the difference between a reactor and a facility, this equates to a UK exposure of 36 reactor-years, per year. My exceedance probability for a major accident in the UK would be 36 times larger, i.e. $p_0 = 7.4 \times 10^{-3}$. The actual UK analysis will be much more nuanced than this, taking into account reactor type and age, and variations in regulations between countries, and possibly taking into

account smaller accidents, other accidents not in the public domain, and near-misses. But this does not stop the final values being anchored on database values.

To give another example from the opposite end of the spectrum, since the start of 2010 (9 years ago at the time of writing) there have been approximately 11 terrorist vehicle-ramming attacks in western European countries, according to the list compiled on the Wikipedia page `https://en.wikipedia.org/wiki/Vehicle-ramming_attack` (downloaded 29 Jan 2019), for which $p_0 = 0.70$ with a 90% 1-year prospective interval of $(0.67, 0.75)$. These values have to be scaled down for the UK, which I believe experiences perhaps one third of all such attacks in Western Europe: for me, 1/3 is a simple resting-place between 1/2, which seems too large, and 1/4, which seems too small. Scaling down in this case is more complicated, because $p_0/3$ is not small. Using a Poisson approach,

$$p_0 = 1 - e^{-\lambda}, \qquad q_0 = 1 - e^{-\lambda\, v} \tag{14a}$$

where $p_0 = 0.70$ is given, q_0 is required, and $v = 1/3$ is the multiplicative factor. Eliminating λ, the solution is

$$q_0 = 1 - (1 - p_0)^v \tag{14b}$$

where the approximation $q_0 \approx p_0\, v$ only holds for small $p_0\, v$. For vehicle ramming attacks in the UK, my beliefs are $q_0 = 0.33$ with a 90% 1-year prospective interval of $(0.31, 0.37)$. Again, the actual UK analysis will be much more nuanced than this, but the final values can still be anchored on database values.

The crucial feature of the above approach is that it starts the SMEs with a well-defined set of tasks, resulting in values which are directly comparable

across hazard classes. For a given hazard class, some of the deviations between database values and final values will be due to deficiencies in the database. These deficiencies can be addressed during the repose period between risk assessments, possiby with science research funding. This type of improvement activity is perfectly aligned with the capabilities of the SMEs, and is the obvious source of changes through time in the risk assessment of a particular hazard class. It should be stressed that the sequence given above does not replace the careful deliberations of SMEs, for which there is still much scope in stages 2 and 5. Instead, its purpose is to shape those deliberations in order to increase their value to the risk manager, particularly in a comparative analysis across hazard classes.

Acknowledgements

This work was supported by the Natural Environment Research Council (NERC) Innovation Internship scheme (grant number NE/P013155/1). I would like to thank the National Risks Team at the Cabinet Office Civil Contingencies Secretariat, for providing a very welcoming and stimulating environment for my Internship. The opinions expressed in this paper are entirely my own.

Appendix

Here is a function to compute $F_{0,k}$ for NUU-HPP, in the statistical computing environment R (R Core Team, 2017).

```
#### compute F_{0,k} for NUU-HPP

## returns the steps of F0 as (x = u, y = Fu) and also p0

## a : the length of the historical time-interval, > 0
## m : the number of events in that interval, >= 0 integer
## k : the length of the prospective period, > 0

F0 <- function(a, m, k = 5) {
  stopifnot(a > 0, m >= 0, m == round(m), k > 0)
  alpha0 <- 0.5 + m
  beta0 <- a; betak <- a + k
  p0 <- 1 - (beta0 / (1 + beta0))^alpha0
  ytop <- qnbinom(0.999, size = alpha0, prob = beta0 / betak)
  y <- seq(from = 0, to = ytop, by = 1)
  u <- 1 - (betak / (1 + betak))^(alpha0 + y)
  list(
    x = c(0, u, 1),
    y = c(0, pnbinom(y, size = alpha0, prob = beta0 / betak), 1),
    p0 = p0)
}

## quantile function, use named arguments in ...

F0inv <- function(p = c(0.05, 0.95), ...) {
  stopifnot(0 <= p, p <= 1)
  ff <- F0(...)
  ii <- findInterval(p, ff$y, left.open = TRUE) # 0 .. N-1
  ff$x[ii + 1L]
}
```

References

Aspinall, W. P. and Cooke, R. M. (2013). Quantifying scientific uncertainty from expert judgment elicitation. In Rougier et al. (2013), chapter 4.

Brown, S. K., Crosweller, H. S., Sparks, R. S. J., Cotterell, E., Deligne, N. I., Guerrero, N. O., Hobbs, L., Kiyosugi, K., Loughlin, S. C., Siebert, L., and Takarada, S. (2014). Characterisation of the Quaternary eruption record: Analysis of the Large Magnitude Explosive Volcanic Eruptions (LaMEVE) database. *Journal of Applied Volcanology*, 3(5). `http://www.appliedvolc.com/content/3/1/5`.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Pacific Grove, CA: Duxbury, 2nd edition.

Cooke, R. M. (2004). The anatomy of a squizzel: The role of operational definitions in representing uncertainty. *Reliability Engineering and System Safety*, 85:313–319.

Cox, D. R. and Donnelly, C. A. (2011). *Principles of Applied Statistics*. Cambridge University Press, Cambridge, UK.

Crosweller, H. S., Arora, B., Brown, S. K., Cotterell, E., Deligne, N. I., Guerrero, N. O., Hobbs, L., Kiyosugi, K., Loughlin, S. C., Lowndes, J., and Nayembil, M. (2012). Global database on Large Magnitude Explosive Volcanic Eruptions (LaMEVE). *Journal of Applied Volcanology*, 1(4). `http://www.appliedvolc.com/content/1/1/4`.

Davison, A. C. (2003). *Statistical Models*. Cambridge University Press, Cambridge, UK.

Ericson, W. A. et al., editors (1981). *The Writings of Leonard Jimmie Savage: A Memorial Selection*. The American Statistical Association and The Institute of Mathematical Statistics.

Goldstein, M. (1997). Prior inferences for posterior judgements. In Chiara, M., Doets, K., Mundici, D., and van Benthem, J., editors, *Structures and Norms in Science. Volume Two of the Tenth International Congress of Logic, Methodology and Philosophy of Science, Florence, August 1995*, pages 55–71. Dordrecht: Kluwer.

Howson, C. and Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach*. Chicago: Open Court Publishing Co., 3rd edition.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin Books Ltd, London, UK.

Kingman, J. F. C. (1993). *Poisson Processes*. Oxford University Press, Oxford, UK.

Klein, G. (2007). Performing a project premortem. *Harvard Business Review*. Available at `https://hbr.org/2007/09/performing-a-project-premortem`.

Lad, F. (1996). *Operational Subjective Statistical Methods*. John Wiley & Sons. Ltd, New York, USA.

Ladyman, J. (2002). *Understanding Philosophy of Science*. Routledge, Abingdon, UK.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., and Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bullentin & Review*, 23(1):103–123.

NRR (2017). National Risk Register (NRR) for Civil Emergencies. Report, UK Cabinet Office. Available at `https://www.gov.uk/government/publications/national-risk-register-of-civil-emergencies-2017-edition`.

OECD (2018). National risk assessments: A cross country perspective. Report, Organisation for Economic Co-operation and Development. Available at `http://dx.doi.org/10.1787/9789264287532-en`.

Perrow, C. (2007). *The Next Catastrophe*. Princeton University Press, Princeton NJ, USA, paperback edition.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, New York, USA.

Rougier, J. C., Sparks, R. S. J., and Cashman, K. V. (2016). Global recording rates for large eruptions. *Journal of Applied Volcanology*, 5:11.

Rougier, J. C., Sparks, R. S. J., and Cashman, K. V. (2018a). Regional volcanism over the last 1000 years. *Journal of Applied Volcanology*, 7:1.

Rougier, J. C., Sparks, R. S. J., Cashman, K. V., and Brown, S. K. (2018b). The global magnitude-frequency relationship for large explosive volcanic eruptions. *Earth and Planetary Science Letters*, 482:621–629.

Rougier, J. C., Sparks, R. S. J., and Hill, L. J., editors (2013). *Risk and Uncertainty Assessment for Natural Hazards*. Cambridge University Press, Cambridge, UK.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66:783–801. In Ericson et al. (1981).

Smith, J. Q. (2010). *Bayesian Decision Analysis: Principle and Practice*. Cambridge University Press, Cambridge, UK.

Sparks, R. S. J., Aspinall, W. P., Crosweller, H. S., and Hincks, T. K. (2013). Risk and uncertainty assessment of volcanic hazards. In Rougier et al. (2013), chapter 11, pages 364–397.

Tetlock, P. E. (2005). *Expert Political Judgment: How good is it? How can we know?* Princeton University Press, Princeton, USA.

Tetlock, P. E. and Gardner, D. (2015). *Superforecasting: The Art & Science of Prediction*. Random House Books, London.

Troffaes, M. C. M. and de Cooman, G. (2014). *Lower Previsions*. John Wiley & Sons, Ltd, Chichester, UK.

van Fraassen, B. (1995). Belief and the problem of Ulysses and the Sirens. *Philosophical Studies*, 77:7–37.

Woo, G. (2015). Understanding the principles of terrorism risk modeling from the Charlie Hebdo attack in Paris. *Defence Against Terrorism Review*, 7(1):33–46.

Woo, G. (2018). Counterfactual disaster risk analysis. *Variance*, 10(2):279–291.